UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This document relates to: Case No. 1:23-cv-08292 Case No. 1:23-cv-10211 Case No. 1:24-cv-00084 Case No. 1:25-cv-03291 Case No. 1:25-cv-03297 Case No. 1:25-cv-03482 Case No. 1:25-cv-03483	Case No. 25-md-3143-SHS-OTW

MEMORANDUM OF LAW IN OPPOSITION TO OPENAI'S MOTION TO STRIKE

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Cases

Other Authorities

Rules

OpenAI is trying to use centralization as an offensive tool to dislodge factual allegations that have been in these cases for years. The Court ordered Plaintiffs to file a consolidated class action complaint (CCAC) with "no new products or causes of action." May 22 Hrg. Tr. 20:15–23, Ex. 3.[1] OpenAI now seeks to twist that directive to its advantage in two ways.

First, OpenAI seeks to expand "no new causes of action" to erase longstanding infringement allegations. Since their inception, Plaintiffs in both the S.D.N.Y. and N.D. Cal. cases have accused OpenAI of infringement by copying Plaintiffs' works from so-called "shadow libraries" on the internet—the modern-day equivalent of Napster for books. *Authors Guild* Compl. ¶ 347, No. 1:23-cv-08292 (S.D.N.Y.), Dkt. 69; *Tremblay* Compl. ¶ 44, No. 3:23-cv-03223 (N.D. Cal.), Dkt. 120; *Alter* Compl. ¶ 92, No. 1:23-cv-10211 (S.D.N.Y.), Dkt. 26; *Chabon* Compl. ¶ 42, No. 3:23-cv-04625 (N.D. Cal.), Dkt. 1; *Silverman* Compl. ¶ 35, No. 3:23-cv-03416 (N.D. Cal.), Dkt. 1.

OpenAI argues that the CCAC is the first complaint to include what it dubs a "downloading claim."[2] But that's wrong. The pre-MDL operative complaint in each class action alleged that OpenAI "obtained" pirated versions of Plaintiffs books from criminal enterprises (like Library Genesis, or "LibGen") that "allow users to download ebooks in bulk." *See, e.g.*, *Authors Guild* Compl. ¶¶ 117, 122, 126. That allegation has been proved true by extensive discovery into the downloading issue. (And that discovery itself confirms that downloading has been in the case all

[1] Unless otherwise noted, all exhibit numbers refer to those exhibits attached to the Walter Declaration, and all "Dkt." citations refer to ECF entries in the instant case, No. 1:25-md-3143.

[2] Calling it a "downloading claim" is a misnomer. The claim is for copyright infringement. Downloading copyrighted books without authorization is a separate act of infringement encompassed by the original copyright-infringement claim. OpenAI's attempt to parse the claim does not make sense because, among other reasons, statutory damages are awarded on a per-work basis. OpenAI is trying to segregate part of the copyright-infringement claim simply to reduce the number of acts of infringement that the factfinder will consider in setting the damages amount.

along.) Former OpenAI employee Ben Mann testified [redacted]. *See infra* at 14–15. OpenAI later tried to delete [redacted] *See* Aug. 12 Hrg. Tr. 160:25–162:2, Ex. 7 (discussing deletion). And Mr. Mann's testimony has been confirmed and expanded upon by the many written discovery requests that Plaintiffs propounded specifically aimed at the downloading question—in response to which OpenAI has provided the requested information. [redacted] Judge Alsup held that this same conduct—by the same person[3]—was "inherently, irredeemably infringing." *Bartz v. Anthropic PBC*, 2025 WL 1741691, at *11 (N.D. Cal. June 23, 2025). Seeing the writing on the wall, OpenAI now seeks to use the MDL as an opportunity to strike allegations that have long been in the case and on which it knows it will almost certainly lose on the merits.

Second, OpenAI seeks to use "no new products" to mean "only the products in the *Authors Guild* action." But the Court permitted products "that have been already asserted in *these cases*," *id.*, and the CCAC's list of products matches *Tremblay*'s. And that makes sense: the MDL should not be used as a vehicle for Defendants to strike claims or allegations that were *already* part of the centralized cases—and being actively litigated—before the MDL was formed. Because there is no serious dispute that the models OpenAI targets were at issue in the operative *Tremblay* complaint before MDL centralization, OpenAI's motion to strike those models should be denied.

OpenAI's gambit is not contemplated by centralization, Rule 12(f), or the Court's orders. Because the CCAC adds no new products or causes of action, OpenAI's motion should be denied.

[3] Ben Mann left OpenAI to found Anthropic, [redacted]

BACKGROUND

Nearly two years after filing, the New York (*Authors Guild*) and California (*Tremblay*) cases[4] were centralized. After the first post-centralization hearing, the Court ordered Plaintiffs to "file a single Consolidated Class Action Complaint . . . which is to include only the same products and causes of action that have already been asserted in the pending putative class actions." Dkt. 60. At the hearing, the Court explained that this approach was meant to sweep in all allegations made in the underlying complaints. *See, e.g.*, May 22 Hrg. Tr. 39:24–25, Ex. 3 ("If there are any models already in the case, that's okay.").

The goal was to avoid relitigating the pleadings. The relevant context, particularly for the "no new causes of action" instruction, was the pending motion to amend in *Tremblay*. *See id.* at 18:13–19, 20:15–23. That motion sought to add new antitrust, breach-of-contract, and other causes of action. *See Tremblay*, Dkt. 370. The Court rejected that motion because it would invite a new round of pleading-based motions: "It seems to me all that would do would be to invite another round of Rule 12(b)(6) motions, and would expand discovery, if you're talking antitrust or additional DMCA claims." May 22 Hrg. Tr. 25:16–19, Ex. 3.

The CCAC follows the Court's order. The CCAC includes the same three copyright causes of action as the *Authors Guild* complaint. *Compare* CCAC ¶¶ 308–327, Dkt. 183, *with Authors Guild* Compl. ¶¶ 412–429, Dkt. 69. And it includes the same products as those alleged in *Tremblay*. *Compare* CCAC ¶ 5, *with Tremblay* Compl. ¶ 36, Dkt. 120.

---

[4] *Authors Guild* and *Tremblay* are themselves consolidated cases. *Authors Guild* consolidated *Authors Guild v. OpenAI, Inc.*, No. 1:23-cv-08292 (S.D.N.Y) and *Alter v. OpenAI, Inc.*, No. 1:23-cv-10211 (S.D.N.Y.). *Tremblay* consolidated *Tremblay v. OpenAI, Inc.*, No. 3:23-cv-03223 (N.D. Cal.); *Silverman v. OpenAI, Inc.*, No. 3:23-cv-03416 (N.D. Cal.); and *Chabon v. OpenAI, Inc.*, No. 3:23-cv-04625 (N.D. Cal.). Plaintiffs refer to the consolidated complaints and the five original complaints as "the underlying complaints."

OpenAI now seeks to narrow the case further. It has ignored the Court's admonition against further "briefing and determination of . . . 12(b)(6) motions," and it instead seeks to "slow everything down" through a motion to dismiss that is thinly veiled as a Rule 12(f) motion to strike. May 22 Hrg. Tr. 25:13–23, Ex. 3. OpenAI has cherry-picked a few changes to isolated paragraphs of the CCAC, saying those changes create a new "downloading claim." OpenAI also says that the relevant models are limited by the New York litigation rather than heeding the Court's guidance to harmonize all the underlying complaints.

OpenAI's motivation is understandable—it read *Bartz*. There, Judge Alsup denied Anthropic's motion for summary judgment as to Anthropic's downloading of pirated books. *See Bartz v. Anthropic PBC*, 2025 WL 1741691 (N.D. Cal. June 23, 2025). *Bartz* did not involve some separate cause of action for "downloading." Instead, the *Bartz* complaint looks like the underlying complaints here: it has the same copyright-infringement cause of action and many of the same (or similar) factual allegations. In his order on Anthropic's fair-use defense, Judge Alsup analyzed Anthropic's downloading and training separately because they are two distinct *uses* of copyrighted material. *Id.* at *11–14. Judge Alsup denied summary judgment as to "copies made from central library copies but not used for training" and found that "downloaded pirated copies used to build a central library were not justified by a fair use." *Id.* at *18–19 (italics omitted).

OpenAI seeks to avoid the same fate. So it has taken a page straight out of Anthropic's playbook. After the summary-judgment order, Anthropic sought a stay pending its application for interlocutory appeal. *Bartz v. Anthropic PBC*, 2025 WL 2308091 (N.D. Cal. Aug. 11, 2025). It claimed that "whether or not it was 'library-building' was never at issue." *Id.* at *2. Judge Alsup rejected the stay and that argument, focusing on the evidence that a download made a separate

copy with a separate use from those copies made for training. *Id.* The Court should do the same here.

Just like Anthropic's, both of OpenAI's arguments on this motion fail. *First*, the "new" downloading claim is not new. Factual allegations are what matter. The underlying complaints' allegations put OpenAI on notice that its infringing activity was at issue, including downloading copyrighted books without authorization. And the same allegations in the CCAC would continue to put OpenAI on notice, even if the motion were granted. *Second*, the California cases included all the models alleged in the CCAC. So OpenAI has "already had a full opportunity" to assert its defenses. May 22 Hrg. Tr. 25:13–23, Ex. 3. Even if these were close questions, the interests of justice and judicial economy favor including and resolving these issues in this case.

LEGAL STANDARD

Under Rule 12(f), "[t]he court may strike . . . any redundant, immaterial, impertinent, or scandalous matter." Motions to strike are disfavored. *See Greenlight Cap., Inc. v. Fishback*, 2024 WL 5168623, at *3 (S.D.N.Y. Dec. 19, 2024). And they are still disfavored even when the argument is that the plaintiff allegedly went beyond the scope of the leave to amend. *See Tian v. Top Food Trading Inc.*, 2023 WL 5200439, at *2 (E.D.N.Y. Aug. 14, 2023); *Rys v. Clinton Cent. Sch. Dist.*, 2022 WL 1541301, at *2 (N.D.N.Y. May 16, 2022). Because motions to strike are ultimately just "tamper[ing]" with the pleadings, there must be "a strong reason for so doing." *Lipsky v. Commonwealth United Corp.*, 551 F.2d 887, 893 (2d Cir. 1976). "Any doubt about whether the challenged material" should be stricken "should be resolved in favor of the non-moving party." 5 Charles A. Wright & Arthur R. Miller, Federal Practice and Procedure § 1382 (3d ed. 2025).

ARGUMENT

I. Allegations that OpenAI illegally downloaded copyrighted books in bulk has long been in the Class Cases

OpenAI moves to strike what it calls the CCAC's "new 'download' claim." Opening Br. at 14, Dkt. 119. At the outset, it is hard to tell what OpenAI thinks is new. Although OpenAI argues against the whole "download theory of infringement," its argument focuses on downloads that were "ultimately *not* used for relevant model training." *Id.* There is no serious dispute that the downloading of books used for training has always been in the case. *See, e.g.*, *id.* ("OpenAI has focused its collection and review of discovery . . . on documents discussing the *compilation, acquisition, and curation* of the text training data *used to train* the relevant models." (first emphasis added)). So OpenAI seems to move against only those allegations that include non-training-related downloads—namely, that works it downloaded from LibGen or other shadow libraries illegally, but may not have been used specifically to train a large language model, are somehow outside the scope of this action.

But all downloading of Plaintiffs' copyrighted books without permission is in the case. *First*, the underlying complaints put OpenAI on notice that unlawful acquisition was at issue. Plaintiffs need only allege facts that amount to infringement. The downloading allegations here easily cleared that bar. Any confusion about these allegations was resolved by consistent and persistent discovery into downloading. And because these allegations would remain, even granting OpenAI's motion would be futile. *Second*, even if there were some gap between the CCAC and the underlying complaints, shifting OpenAI's downloading to an entirely new case—on a separate track from the case involving the *same* downloading once those downloads are used for training—would waste time, resources, and run counter to the "public interest in these things being resolved." May 22 Hrg. Tr. 36:20–24, Ex. 3.

A. Downloading has been in the case at every stage

Plaintiffs' argument that OpenAI infringed by downloading terabytes of pirated material is not new. Downloading allegations (including downloads ultimately not used for training) are in all the underlying complaints. Any ambiguity about the scope of these allegations has been resolved in discovery. And because OpenAI's motion does not seek to strike all these factual allegations or somehow claw back all the proof of them found through discovery, even granting its motion would not cleanse this case of the substance of OpenAI's unauthorized downloading.

1. The Complaints have long alleged OpenAI infringed through unlawful acquisition of pirated copies of books

Downloading works from illegal pirate websites has always been alleged as a basis for infringement. Tellingly, OpenAI does not move to strike any downloading-related factual allegations. Instead, it protests that the CCAC reveals a new "download *theory* of infringement." Opening Br. at 14, Dkt. 119 (emphasis added). That protest is far afield from the Court's instruction not to add any new "causes of action" while discussing potential "antitrust or additional DMCA claims." May 22 Hrg. Tr. 25:16–19, Ex. 3. But the argument also fails on its merits. It is enough that the underlying complaints alleged infringing acts of downloading. Plaintiffs were under no obligation to plead the legal theory explaining why downloading was an act of infringement.

Under Rule 8, "a plaintiff is not required to plead every legal theory in support of a claim. The pleading need only provide the defendant with fair notice of the claims alleged." *Greicus v. Liz Claiborne, Inc.*, 2002 WL 244598, at *3 (S.D.N.Y. Feb. 20, 2002) (Stein, J.) (citations omitted).[5] "Factual allegations alone are what matters." *Albert v. Carovano*, 851 F.2d 561, 571 n.3

[5] *See also Johnson v. City of Shelby*, 574 U.S. 10, 11 (2014) ("Federal pleading rules call for 'a short and plain statement of the claim showing that the pleader is entitled to relief'; they do not countenance dismissal of a complaint for imperfect statement of the legal theory supporting the claim asserted." (quoting Fed. R. Civ. P. 8(a)(2)); *Skinner v. Switzer*, 562 U.S. 521, 530 (2011)

(2d Cir. 1988). For a copyright claim, "a complaint must allege" (1) the works at issue, (2) ownership, (3) registration, and (4) "by what acts during what time the defendant infringed the copyright." *Cobb v. Am. Urb. Radio Networks LLC*, WL 641437, at *2 (S.D.N.Y. Feb. 27, 2025). "The fourth prong requires . . . some factual allegations to narrow the infringing acts beyond broad conclusory statements of infringement." *Id.* (citation omitted).

The underlying complaints have dozens of allegations describing OpenAI's infringing downloads. These allegations are collected in Table 1, in the Appendix. These allegations describe both training-related and non-training-related downloads. *See, e.g.*, *Authors Guild* ¶ 100 ("OpenAI has admitted that 'training' LLMs 'require[s] large amounts of data,' and that 'analyzing large corpora' of data '*necessarily involves first making copies of the data to be analyzed*.'" (emphasis added)); *Authors Guild* ¶ 415 ("Defendants infringed . . . by, *among other things*, reproducing the works. . . in datasets used to train their artificial intelligence models." (emphasis added)); *Tremblay* ¶ 58 ("[W]hether Defendants violated the copyrights of Plaintiffs and the Class *when they downloaded copies* of Plaintiffs' copyrighted books *and* used them to train ChatGPT." (emphasis added)); *Tremblay* ¶ 71 ("Plaintiffs never authorized OpenAI to make copies of their books[.]"); *Alter* ¶ 80 ("Millions of copyrighted works were *copied*—including at least tens of thousands of nonfiction books—*and then ingested* for the purpose of 'training' Defendants' GPT models."

("[A] complaint need not pin plaintiff's claim for relief to a precise legal theory. Rule 8(a)(2) . . . generally requires only a plausible 'short and plain' statement of the plaintiff's claim, not an exposition of his legal argument."); *Baker v. Latham Sparrowbush Assocs.*, 808 F. Supp. 981, 989 (S.D.N.Y. 1992) ("Federal pleading is by statement of claim upon which relief may be based rather than by legal theory. Plaintiffs need only plead facts showing entitlement to relief and are not required to specify the legal theory upon which the claim is based." (citation omitted)), *aff'd*, 72 F.3d 246 (2d Cir. 1995); *Davis v. Rumsey Hall Sch., Inc.*, 2023 WL 6379305, at *9 (D. Conn. Sept. 29, 2023) ("Plaintiff need not broadcast its strategy at the pleading stage, and need not spell out each legal theory on which he might rely to prove his claims, when varying theories can be supported by his factual allegations.").

(emphasis added)); *Alter* ¶ 87 ("OpenAI . . . chose to copy a massive corpus of copyrighted books right from the internet, almost certainly from illegal sources . . . without even paying for an initial copy."); *Chabon* ¶ 34 (OpenAI's "practice *necessarily leads OpenAI to capture, download, and copy copyrighted written works*, plays and articles" (emphasis added)); *Chabon* ¶ 60 ("Whether Defendants violated the copyrights of Plaintiffs and the Class when they downloaded and copied Plaintiffs' and the Class's copyrighted books").

These allegations put OpenAI on notice of the "acts" by which it "infringed the copyright[s]." Plaintiffs were never required to spell out the idea that the downloading allegations were an independent way to prove their direct-infringement claim. It is enough that (1) Plaintiffs *alleged the fact* that OpenAI downloaded a trove of pirated material without permission and (2) downloading pirated material is a valid *legal theory* of infringement. *See Oneida Indian Nation v. County of Oneida*, 617 F.3d 114, 139 (2d Cir. 2010) ("[T]he essence of a cause of action is found in the facts alleged and proven by the plaintiff, not the particular legal theories articulated.").

OpenAI makes three arguments, but none moves the needle.

First, OpenAI says many allegations discuss downloads alongside training, so only training-related downloads are included. But that is not always the case, as shown above and in Table 1. And in any event, using training to explain OpenAI's motivation or to provide context for the downloads is not the same as *limiting* the factual description of OpenAI's conduct. *See, e.g.*, Opening Br. at 13 (citing *Tremblay* ¶ 65 ("[T]o train the OpenAI Language Models, OpenAI relied on harvesting mass quantities of textual material from the public internet, including Plaintiffs' books, which are available in digital formats.")). The allegations about OpenAI's conduct control, and the Court should not turn Rule 12 on its head by construing Plaintiffs' allegations in the light most favorable to *OpenAI*. *See EEOC v. Kelley Drye & Warren, LLP*, 2011 WL 3163443, at *2

(S.D.N.Y. July 25, 2011) ("When reviewing a motion to strike, the court views the pleading under attack most favorably to the pleader." (citation omitted)).

Second, OpenAI's motion isolates just a few paragraphs that have been rephrased and labels them "new." But that is the wrong way to read the complaint. "[T]he factual allegations in a complaint" must be read "as a whole," and the allegations listed above show that downloading was always included. *Koury v. Xcellence, Inc.*, 649 F. Supp. 2d 127, 133 (S.D.N.Y. 2009) (quoting *Shapiro v. Cantor*, 123 F.3d 717, 719 (2d Cir. 1997)).

And even if OpenAI's cherry-picked paragraphs were considered on their own, they are not new. Their language is grounded in the underlying complaints (as shown in Table 2, in the Appendix). *Compare, e.g.*, CCAC ¶ 302 ("Defendants have . . . reproduc[ed] and maintain[ed] the Class's works without consent and or compensation, and in Defendants' LLM 'training.'"), *with, e.g.*, *Authors Guild* ¶ 424 (alleging that Microsoft built the "system that OpenAI used to maintain and copy the copyrighted works"); *Tremblay* ¶ 39 ("[Novels] were copied into the BooksCorpus dataset without consent, credit, or compensation to the authors.").

Similarly, the supposedly new class definition is pulled from the underlying complaints' substantive allegations, so it does not change the scope of discovery. *Compare* CCAC ¶ 294 (using the phrase "downloaded or otherwise reproduced"), *with Alter* ¶ 115(a) ("Whether Defendants' reproduction of the Class's copyrighted work constituted a copyright infringement."); *Chabon* ¶ 60 ("Whether Defendants violated the copyrights of Plaintiffs and the Class when they downloaded and copied Plaintiffs' and the Class's copyrighted books."). And even if the class definition represents some evolution to conform it with the substantive allegations and evidence, that change is inevitable and routinely permitted. *See Woe ex rel. Woe v. Cuomo*, 729 F.2d 96, 107 (2d Cir. 1984) ("It is often proper . . . for a district court to view a class action liberally in the early stages

of litigation, since the class can always be modified or subdivided as issues are refined for trial."). Here, Plaintiffs are just getting ahead of the curve by borrowing aspects of the class definition that Judge Alsup certified in *Bartz*. *See Bartz v. Anthropic PBC*, 2025 WL 1993577, at *6 (N.D. Cal. July 17, 2025) (certifying class of "[a]ll beneficial or legal copyright owners of the exclusive right to reproduce copies of any book in the versions of LibGen or PiLiMi downloaded by Anthropic.").

Third, OpenAI claims that the Court already rejected downloading by denying the *Tremblay* motion to amend. Not so. As discussed at the May 22 hearing, the *Tremblay* amendments involved new causes of action, including antitrust and DMCA claims. In fact, the only time OpenAI even mentioned downloading in its brief opposing the *Tremblay* motion to amend was to note that downloading was *already in the case*: "Plaintiffs alleged in the FCAC that OpenAI 'downloaded copies of Plaintiffs' copyrighted books.' FCAC ¶ 58(a)." *Tremblay*, Dkt. 401 at 15.

At bottom, OpenAI's claim that it was not on notice is not credible. OpenAI should not be blindsided by the idea that its downloading billions of pages of pirated books is part of the copyright case against it. In *Bartz*, "Anthropic's downloading of pirate libraries and its deployment of bit-torrenting to do so looms large." *Bartz*, 2025 WL 2308091, at *1. So too here. Table 3 in the Appendix compares the complaints from *Bartz* and *Alter* (one of the underlying complaints here). In those two complaints, the claims for relief and the typicality, commonality, and predominance allegations are essentially identical. The complaints are even structured similarly. The class definitions vary just slightly, with *Bartz*'s definition including material "used by Defendant in LLM training, research, or development" while *Alter*'s includes material "used by Defendants in training their generative artificial intelligence models." But if OpenAI's argument turns on including "research or development" in the complaint's *proposed* class definition, that just shows the thinness of its position. Federal pleading does not permit (let alone encourage) this kind of

paragraph-parsing reading of the complaint, and Plaintiffs and the Court can modify a class definition at any later stage, including when class certification is briefed and after a class is certified. *Dornberger v. Metro. Life Ins. Co.*, 182 F.R.D. 72, 74 (S.D.N.Y. 1998) ("[T]he definition of the class is conditional and may be modified by the Court at any time."); *Andres v. Town of Wheatfield*, 621 F. Supp. 3d 415, 419 (W.D.N.Y. 2022) ("Courts have recognized the ongoing refinement and give-and-take inherent in class action litigation, particularly in the formation of a workable class definition and have allowed Plaintiffs to amend a class definition even in their reply brief at the certification stage." (citation omitted)).

Indeed, "[o]ne of the most important objectives of the federal rules is that lawsuits should be determined on their merits and according to the dictates of justice, rather than in terms of whether or not the averments in the paper pleadings have been artfully drawn." *Koury*, 649 F. Supp. 2d at 133 (quoting Wright & Miller § 1286). The Court must read each complaint as a whole and "construe the pleading in the [plaintiff's] favor, whenever the interest of justice so requires." Wright & Miller § 1286.

Thus, if there was any ambiguity about whether the complaints provided adequate notice, that ambiguity should be resolved in Plaintiffs' favor. That is especially so when discovery has provided additional notice, and the discovery taken so far mitigates OpenAI's prejudice argument.

2. Discovery confirms that OpenAI's decision to download books from the internet was at issue in the Class Cases before the MDL

Because Rule 8 requires only a short and plain statement, discovery often sharpens the issues and clarifies how the pleadings, claims, and evidence will fit together. *Anvik Corp. v. Samsung Elecs.*, 2009 WL 10695623, at *2 (S.D.N.Y. Sept. 16, 2009) ("The generous interpretation of Rule 26 discovery also complements the nature of federal civil pleadings by

allowing a short, plain statement of a claim to support a cause of action and then permitting the parties to engage in discovery to develop the facts, theories, and defenses of the case.").[6]

Just so here. Even if the complaints were not crystal clear, the discovery process sharpened their focus. That sharpening put Defendants on notice, starting more than a year ago. And Defendants' behavior reveals that they were in fact on notice of Plaintiffs' allegation that OpenAI's piracy itself constitutes copyright infringement.

First, as far back as eighteen months ago, Plaintiffs have served substantial written discovery aimed at OpenAI's downloads of copyrighted material from pirate websites. Table 4 in the Appendix collects these many requests, which cover material not used for training. *See, e.g.*, RFP 109 ("Documents sufficient to show all books You have downloaded and the source of such books."); Lepic Dec. 26 Email, Ex. 11. For example, in March, Plaintiffs served an interrogatory asking OpenAI to "[i]dentify all datasets in [its] possession that contain copyrighted works." Third Set of ROGs, ROG 21, Ex. 9. And "dataset" was defined to include all downloads, "even if not ultimately used for training." *Id.* at Def'n 3. After negotiating for months, OpenAI agreed to [redacted] And it [redacted] Gorman July 29 Email ¶ 4, Ex. 12. Granted, OpenAI of course objected. But it did not categorically object

[6] *See also Fletcher v. Atex, Inc.*, 1993 WL 97321, at *1 (S.D.N.Y. Mar. 30, 1993) ("The precise contours of the claims can certainly be established in the discovery process."); *Gen. Elec. Co. v. Bucyrus-Erie Co.*, 563 F. Supp. 970, 977 (S.D.N.Y. 1983) ("All that is required is that the claim for relief give notice to the opposing side. Liberal discovery rules allow 'whatever additional sharpening of the issues may be necessary.'" (quoting *George C. Frey Ready-Mixed Concrete, Inc. v. Pine Hill Concrete Mix Corp.*, 554 F.2d 551, 554 (2d Cir. 1977)).

to all downloading-related material (as opposed to downloads not used for training). *See* Opening Br. at 14 (recognizing that discovery has focused on "compilation [and] acquisition" of training data). And the parties have been haggling over non-training-related downloads for much of discovery. So any claim of lack of notice is implausible.

Second, the deposition of Ben Mann, [redacted]

[7] Books1 and Books2 are book-based datasets that OpenAI used to train its models. CCAC ¶ 113. "[T]hese datasets were sourced and downloaded from the notorious pirate website LibGen." *Id.*

[redacted]

Third, downloading has also been discussed at hearings. Back in November 2023, Plaintiffs stated the copyright claim straightforwardly and broadly: "The violation of the copyright law is the willful copying and reproduction of his books in their entirety." Nov. 29 Hrg. Tr. 11:15–17, Ex. 1. The first reproduction is the download.

More recently, at the March 27 JPML hearing, Plaintiffs noted that "there's use issues about illegal acquisition, there's use issues about training." JPML Hrg. Tr. 28:7–10, Ex. 2. The panel decided to centralize these cases on this "limited number of buckets." *Id.* at 28:11–12.

At the May 22 hearing with the Court, Plaintiffs noted that "we have two theories of infringement. The first is the piracy that occurred. And under *Andy Warhol*, each use is a separate infringement. That piracy, that initial download, we have the facts for that. . . . So with literally a couple of months more of discovery, maybe with a few expert reports, we can certainly get to a

resolution of the piracy use of that. Now, there's a separate issue with respect to training." May 22 Hrg. Tr. 51:25–52:9, Ex. 3.

Similarly, at the June 26 tech tutorial, Plaintiffs said: "The allegations in this case are that OpenAI and Microsoft downloaded and built digital libraries made out of not just one book or one news article, but millions of books and millions of articles without permission. As to books, OpenAI and Microsoft purposely sought out and sourced these books from, among other sources, known pirated websites, including sites like Library Genesis." June 26 Hrg. Tr. 152:5–11, Ex. 6; *see also id.* at 158:9–159:14 (discussing data acquisition by pirating).

And at the May 27 discovery hearing before Judge Wang, Plaintiffs' counsel noted that "we have a direct infringement claim against Microsoft that we have alleged since the beginning. So to the extent they downloaded Library Genesis . . . that is squarely within the case." May 27 Hrg. Tr. 25:6–20, Ex. 4; *see also id.* at 77:18–78:6 (similar, noting that acquisition alone goes to knowledge and willfulness).

Even Defendants have recognized the role of downloading. At the June 25 discovery hearing, Microsoft acknowledged that at least training-related downloads are at issue: "The direct infringement alleged in this case is the downloading and use of copyrighted materials from pirated sites for training. I mean, their own complaint defines that scope of relevance." June 25 Hrg. Tr. 12:24–13:2, Ex. 5. And as recently as the August 12 discovery hearing, OpenAI said that "there's been extensive discovery" about downloading, including "testimony about when the books were downloaded, who was involved in that downloading, how it was downloaded." Aug. 12 Hrg. Tr. 165:15–23, Ex. 7.

Rather than a legitimate complaint regarding the scope of discovery, OpenAI's motion is a longshot attempt to slice off downloading precisely *because* the discovery conducted to date

regarding OpenAI's downloading *has been so damning*. Discovery has revealed that OpenAI downloaded [redacted] from LibGen. [redacted] It did so despite LibGen's having been repeatedly enjoined by courts in this district for copyright infringement. *See, e.g., Elsevier Inc. v. Sci-Hub*, 2017 WL 3868800 (S.D.N.Y. June 21, 2017). And it hid this conduct from the public [redacted] The Court should not bless OpenAI's attempt to hide its massive downloading of pirated books from judicial scrutiny. Given both the complaints' allegations and the course of discovery, OpenAI has been on notice that Plaintiffs are asserting a theory of copyright infringement that easily encompasses OpenAI's downloading of Plaintiffs' copyrighted works without permission. The Court should see OpenAI's attempt to rewrite history for what it is.

3. Granting the motion would be futile

For at least two reasons, even granting OpenAI's motion would not get it where it wants to go. These reasons go hand in hand with the above discussion of allegations versus legal theories.

First, OpenAI's motion does not seek to strike every allegation about downloading (likely because it knows that those allegations are in the underlying complaints). Instead, it targets phrasing changes in three isolated paragraphs of the CCAC. But those paragraphs do not represent the entire downloading "claim."[8] Even if the Court struck those paragraphs, there would still be many paragraphs describing the substance of OpenAI's piracy. *See, e.g.*, CCAC ¶¶ 110–121, 166–169, 182, 279. Those substantive allegations control, so downloading would still be in the case.

[8] Plaintiffs recognize the awkwardness in determining how many paragraphs make up a claim. But that Sorites-paradox problem is just more proof that OpenAI's motion is an improper attempt to get a substantive win. *See Day v. Moscow*, 955 F.2d 807, 811 (2d Cir. 1992) (motions to strike are not "designed for . . . dismissal of claims in their entirety"); *Koch v. Dwyer*, 2000 WL 1458803, at *1 (S.D.N.Y. Sept. 29, 2000) (motions to strike are "not an appropriate vehicle to dismiss claims from a complaint").

See supra at 5 & n.2; *see also Samuel v. Rose's Stores, Inc.*, 907 F. Supp. 159, 162 (E.D. Va. 1995) (denying motion to strike allegations after "Plaintiff did exceed the scope of the leave to amend . . . however, the changes in the complaint do not affect the substance of the claims against the Defendant.").

Second, OpenAI's motion reflects "a fundamental misunderstanding of the need to amend pleadings." *McCree v. City of New York*, 2023 WL 1825184, at *2 (E.D.N.Y. Feb. 8, 2023). Even if, for instance, the underlying complaints included allegations about only those downloads used for training, Plaintiffs do not need to amend to include all downloads. "Amended pleadings are not necessary simply to update a complaint to reflect those facts revealed in discovery—a party may rely on such facts in summary judgment or trial, even if not in the operative pleading." *Id.* "Here there is no indication that the proposed new facts deviate so substantially from the prior facts alleged; they are simply additional support for existing claims." *Id.* Rather, all downloading is closely related to downloading for training and training itself. *See Bartz*, 2025 WL 2308091, at *1 ("Furthermore, Anthropic recopied only *some* 'subsets' of the pirated works to use for training large language models, or LLMs, but kept them *all*. Which works in its collection were actually used to train LLMs, which were not, and why were all retained? Anthropic has refused to come clean on this, even now, and for all we know, most were never used (or not solely used) to train LLMs." (emphases in original)). So striking allegations about downloading would not change the scope of the case when OpenAI's piracy has been revealed in discovery and provides further evidence supporting the original copyright-infringement claim.

B. Granting the motion would waste party and judicial resources

Even if it were a close call or if downloading were not in the underlying complaints, it should still be included now. Getting rid of downloading would not significantly limit discovery

in this case. But it would be hugely inefficient to shove this theory of infringement onto another case.

1. Granting the motion would not meaningfully limit discovery

Granting the motion would not streamline the case, so there is no "strong reason" to strike these allegations. *Lipsky*, 551 F.2d at 893.

First, downloading has already been a part of discovery. As discussed above, many of Plaintiffs' discovery requests have been aimed at downloading in general, and OpenAI has been providing information about its downloading. The simplest explanation for that conduct is that downloading *has always been in the case*. Indeed, OpenAI has provided discovery into exactly what it claims is out of the case: the initial downloading [redacted] of books from LibGen. [redacted] OpenAI has even purported to produce [redacted] Indeed, Plaintiffs' most recent interrogatories asked for "the names of each dataset that OpenAI *downloaded* from a known shadow library." OpenAI's Suppl. R&Os to Third Set of ROGs, Ex. 10 (emphasis added). And as mentioned, in response to these interrogatories, OpenAI investigated [redacted] Gorman July 29 Email ¶ 4, Ex. 12. Investigation into OpenAI's downloading of pirated materials has substantially progressed.

Second, discovery into OpenAI's willful and bad faith copying will necessarily encompass discovery into all downloads. Willfulness is essentially a "totality of the circumstances" test. *Almond Int'l, Inc. v. Arpas Int'l Ltd.*, 1999 WL 476287, at *1 (S.D.N.Y. July 8, 1999). It "may be

inferred from the defendant's conduct," and "courts look to several factors," including notice that the work was protected, warnings, and experience with copyright. *Agence France Presse v. Morel*, 2014 WL 3963124, at *3 (S.D.N.Y. Aug. 13, 2014) (citation omitted). And under the first fair-use factor, courts must consider bad faith, such as whether "defendants must have known" that a work "was acquired in an unauthorized fashion." *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 477 (2d Cir. 2004); *see also Rogers v. Koons*, 960 F.2d 301, 309 (2d Cir. 1992) (considering infringer's removing the copyright notice as part of the fair-use analysis).

Here, OpenAI's entire course of conduct reveals its willfulness and bad faith. Even if the copyright claim were limited to training-related downloads or no downloads at all, all downloads would be relevant because they reveal the sheer scale and regularity of OpenAI's piracy (not to mention the illicit sources). *See Kadrey v. Meta Platforms, Inc.*, 2025 WL 1752484, at *11 (N.D. Cal. June 25, 2025) ("Meta's use of shadow libraries is relevant to the issue of bad faith[.]"); *UMG Recordings, Inc. v. Escape Media Grp., Inc.*, 2015 WL 1873098, at *4 (S.D.N.Y. Apr. 23, 2015) (admitting evidence of the degree of willfulness even after the infringement was found to be willful). Similarly, OpenAI's process for filtering the downloaded files to create training data is relevant. That will require an investigation of the downloaded files and a comparison to the resulting training data. For instance, whether and how OpenAI systematically removed copyright notices from all its downloads shows its willful disregard for copyright protections even if the claim were limited. *Rogers*, 960 F.2d at 309.[9] Because this information is relevant regardless, granting the motion would not significantly limit the scope of discovery.

[9] Differences between the downloaded files and the resulting training data would also be relevant under fair use factor three (the amount and substantiality of the portion used).

2. Requiring a separate downloading-only case would be inefficient

As noted above, downloading is in this case whether OpenAI succeeds on this motion or not. But requiring a new case to start from scratch would create a far greater burden—for both plaintiffs *and* OpenAI. OpenAI's burden argument rings hollow when it will have to produce the exact same data in another case, but without the head start of the requests, investigations, and productions already made here.

Even if the Court found that a downloading claim was not technically part of the underlying complaints, the federal rules discourage wasting judicial and party resources in just this situation. *See Bytemark, Inc. v. Xerox Corp.*, 2022 WL 94859, at *12 (S.D.N.Y. Jan. 10, 2022) ("Defendants have been on notice of these claims, whose underlying factual allegations are already a part of the previous complaints, and therefore will not be unduly prejudiced by having to litigate them. Indeed, when a plaintiff's proposed new claims arise out of the same facts set forth in the original complaint, forcing plaintiffs to institute a new action against the defendant would run counter to the interests of judicial economy." (cleaned up)).

Indeed, even as late as summary judgment, courts will permit "claims that are related to or are mere variations of previously pleaded claims—that is, claims based on the same nucleus of operative facts and similar legal theories as the original claims." *Coudert v. Janney Montgomery Scott, LLC*, 2005 WL 1563325, at *3 (D. Conn. July 1, 2005) (permitting a hostile-work-environment theory of employment discrimination in a brief opposing summary judgment when the complaint explicitly pleaded only disparate treatment and retaliation), *aff'd*, 171 F. App'x 881 (2d Cir. 2006); *see also Hanlin v. Mitchelson*, 794 F.2d 834, 841 (2d Cir. 1986) (permitting amendment at summary judgment to add negligence and contract claims because they were

"merely variations on the original theme of malpractice, arising from the same set of operative facts as the original complaint").

The preference for avoiding waste and deciding cases on the merits is at its apex here. The relevant facts were alleged in the underlying complaints. Discovery into downloading is already well underway. Another court has already endorsed precisely this theory of liability. And OpenAI is trying to create an entirely new case for downloads that it *never used* when its principal defense is *fair use*. The Court should not allow OpenAI's gamesmanship to eliminate the most straightforward theory of liability and waste years more of the parties' and the Court's time.

* * *

OpenAI does not argue that the CCAC makes *any* new factual allegations. Instead, OpenAI says the CCAC presents a new "theory." But Plaintiffs are under no obligation to plead "theories" at all. Instead, Rule 8 requires notice of the *acts* by which OpenAI infringed. The underlying complaints and the parties' conduct throughout discovery have provided ample notice to OpenAI. Further, striking a few isolated paragraphs would not limit Plaintiffs' ability to seek discovery and present downloading evidence supporting their copyright claim. Finally, even if it were a close question, every tiebreaker favors Plaintiffs: Rule 8 favors liberal construction and reading the complaint as a whole; motions to strike are disfavored, especially when cooked-up to work a substantive change in pre-existing allegations; and the interests of justice and judicial economy favor resolving disputes on their merits and in a single case. So long as there is "[a]ny doubt about whether the challenged material" should be stricken, the motion "should be resolved in favor of" Plaintiffs. Wright & Miller § 1382.

II. The models should not be stricken

OpenAI has also been on notice that the CCAC's named models are in the case.[10] Although Judge Wang limited discovery to those models identified in OpenAI's interrogatory responses, even OpenAI acknowledges that at least one other model is included. *See* Opening Br. at 6 n.15, Dkt. 119. And the Court's order did not refer to the interrogatory response but instead said that the CCAC should be limited to "the same products . . . that have already been *asserted* in the pending putative class actions." Dkt. 60 (emphasis added); *see also* May 22 Hrg. Tr. 39:24–25, Ex. 3 ("If there are any models already in the case, that's okay.").

Tremblay has always been broader than OpenAI's interrogatory response in *Authors Guild*. *First*, the consolidated *Tremblay* complaint alleged that "OpenAI has made other language-model variants that are in commercial use but are not publicly accessible. . . . OpenAI CEO Sam Altman confirmed that GPT-5 is under development. Together, OpenAI's large language models, including any in development, will be referred to as the 'OpenAI Language Models.'" ¶ 36. It also included "variant forms," like "gpt-4-0125-preview, gpt-4-turbo-preview, and gpt-4-32k." *Id.* (internal quotation marks omitted). *Tremblay* thus included versions that were not yet publicly available, like GPT-4V and GPT-4.5, as well as GPT-5 and any other models "in development." This allegation matches the CCAC's list of models. CCAC ¶ 5.

Second, the discovery in *Tremblay* involved these models. The *Tremblay* court ordered OpenAI to produce "the text pre-training data for in-development, text-based GPT-class of models . . . including the next GPT-class model still being developed, which has been referred to as 'Orion.'" *Tremblay*, Dkt. 247 at 4. [redacted] The court also

[10] OpenAI agrees that GPT-3, GPT-3.5, GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, GPT-4o, and GPT-4o Mini are in the case. Opening Br. at 6. The parties dispute whether GPT-4V(ision), GPT-4.5, GPT-5, and derivative and successor models are in. *Id.*

ordered that, in responding to other document requests, OpenAI "*not exclude* documents because they concern a GPT-class model in development." *Id.* at 3–4 (emphasis in original). So OpenAI was already ordered to produce documents about these models more than seven months ago.

Plus, the court found that discovery about the models in development would be relevant to issues like fair use for the existing models. *See Tremblay* Dec. 17 Hrg. Tr. 48:20–52:6, Ex. 8. So even striking the models from the CCAC would not narrow discovery.

Third, the interests of justice and efficiency again favor including these models. In granting injunctions against infringers, courts have reasoned that "[r]equiring Plaintiff to commence litigation for each future violation would be an extreme hardship, while preventing Defendant from continually infringing on Plaintiffs' copyrighted material is not." *Eileen Grays, LLC v. Remix Lighting, Inc.*, 2019 WL 6609834, at *4 (N.D.N.Y. Dec. 5, 2019). The same logic applies here. OpenAI does not argue that any of the legal issues are different for these models. So filing a new case for each one would be a formality. The Court should not require Plaintiffs to play whack-a-model.

CONCLUSION

For these reasons, OpenAI's motion at Dkt. 118 should be denied.

Dated: August 14, 2025
New York, New York

Respectfully submitted,

/s/ Justin Nelson

Justin A. Nelson (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Telephone: 713.651.9366
jnelson@susmangodfrey.com

Interim Lead Counsel for Class Plaintiffs

Alejandra C. Salinas (*pro hac vice*)
Amber B. Magee (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Tel.: 713.651.9366
asalinas@susmangodfrey.com
amagee@susmangodfrey.com

Rohit D. Nath (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067
Tel.: 310.789.3100
rnath@susmangodfrey.com

J. Craig Smyser
Charlotte Lepic
Henry Walter
SUSMAN GODFREY L.L.P.
One Manhattan West, 51st Floor
New York, NY 10001
Tel.: 212.336.8330
csmyser@susmangodfrey.com
clepic@susmangodfrey.com
hwalter@susmangodfrey.com

Jordan W. Connors (*pro hac vice*)
SUSMAN GODFREY L.L.P.
401 Union Street, Suite 3000
Seattle, WA 98101
Tel.: 206.516.3880
jconnors@susmangodfrey.com

/s/ Joshua Michelangelo Stein
Joshua Michelangelo Stein (California SBN 298856)
Maxwell V. Pritt (pro hac vice anticipated)
BOIES SCHILLER FLEXNER LLP
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
(415) 293-6800
jstein@bsfllp.com
mpritt@bsfllp.com

Counsel for Class Plaintiffs

APPENDIX

TABLE 1: DOWNLOADING ALLEGATIONS

Authors Guild (No. 1:23-cv-08292 (S.D.N.Y.), Dkt. 69, all emphasis added)	• "OpenAI has admitted that 'training' LLMs 'require[s] large amounts of data,' and that 'analyzing large corpora' of data '*necessarily involves first making copies of the data to be analyzed*.'" ¶ 100. • "The similarities in the sizes of Books2 and Books3, and the fact that there are only a few pirate repositories on the Internet that allow bulk ebook downloads, *strongly indicates that the books contained in Books2 were also obtained from one of the notorious repositories discussed above*." ¶ 122. • "In short, OpenAI admits it needs and uses 'large, publicly available datasets that include copyrighted works'—and specifically, 'high-quality' copyrighted books—to 'train' its LLMs; pirated sources of such 'training' data are readily available; and one or more of these sources contain Plaintiffs' works." ¶ 127. • "As the public company made its decision to invest $13 billion into OpenAI, surely Microsoft—like Andreesen Horowitz—was fully aware that *OpenAI was taking a massive corpus of copyrighted content, without compensation to rightsholders, and copying it for the purpose of training and developing its GPT models* to mimic the human writing. ¶ 138. • "There are questions of fact or law common to the Classes, including: a. Whether Defendants copied works owned by Plaintiffs and the members of the Classes[.]" ¶ 405.
Tremblay (No. 3:23-cv-03223 (N.D. Cal.), Dkt. 120, all emphasis added)	• "On information and belief, the reason ChatGPT can accurately summarize a certain copyrighted book is because that book was copied by OpenAI *and* ingested by the underlying OpenAI Language Model (either GPT-3.5 or GPT-4) as part of its training data." ¶ 50. • "Numerous questions of law or fact common to each Class arise from Defendants' conduct: a. whether Defendants violated the copyrights of Plaintiffs and the Class *when they downloaded copies* of Plaintiffs' copyrighted books *and* used them to train ChatGPT" ¶ 58. • "Plaintiffs *never authorized OpenAI to make copies of their books*, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works). All those rights belong exclusively to Plaintiffs under copyright law" ¶ 71.

Alter (No. 23-cv-10211 (S.D.N.Y.), Dkt. 26, all emphasis added)	• "That data, as explained in more detail below, consists largely of copyrighted material. . . . ChatGPT could only be trained to generate this range of expression by making *unlicensed reproductions of a massive corpus of copyrighted content*, including Plaintiffs' works and works owned by the Class. Without this largescale infringement—without a large corpus of copyrighted material from which to mine expression—Defendants' GPT models could not have been trained to perform their intended function, that is, to mimic human written expression." ¶ 54. • "Millions of copyrighted works were *copied*—including at least tens of thousands of nonfiction books—*and then ingested* for the purpose of 'training' Defendants' GPT models. Those works were used as inputs into the GPT models, *then copied many times again* to gauge how well the output mimicked human expression[.]" ¶ 80. • "OpenAI and Microsoft have also deprived authors of books sales and licensing revenues. There is, and has been, an established market for the sale of books and e-books, yet OpenAI ignored it and *chose to copy a massive corpus of copyrighted books right from the internet*, almost certainly from illegal sources . . . without even paying for an initial copy." ¶ 87.
Chabon (No. 3:23-cv-04625 (N.D. Cal.), Dkt. 1, all emphasis added)	• "While casting a wide net across the internet to capture the most comprehensive set of content available allows OpenAI to better train its GPT models, this practice *necessarily leads OpenAI to capture, download, and copy copyrighted written works*, plays and articles." ¶ 34. • "Among the content OpenAI has scraped from the internet to construct its training datasets are Plaintiffs' copyrighted works." ¶ 35. • "ChatGPT can . . . summarize a certain copyrighted book and provide in-depth analysis of that book is because it was copied by OpenAI *and* copied and analyzed by the underlying GPT model as part of its training data." ¶ 47. • "This action involves common questions of law and fact, . . . including . . . b. Whether Defendants violated the copyrights of Plaintiffs and the Class *when they downloaded and copied Plaintiffs' and the Class's copyrighted books*[.]" ¶ 60.

TABLE 2: OPENAI'S THREE PARAGRAPHS	
CCAC (Dkt. 183)	Underlying Complaints
Nature of the Case (¶ 14) "Plaintiffs seek to represent a Class of book copyright holders whose works were used by Defendants in conjunction with their artificial intelligence models."	*Alter*: "OpenAI and Microsoft created unlicensed reproductions of the copyrighted works . . . *in the course of* training and fine-tuning their models." ¶ 71 (emphasis added); *see also* ¶ 80 ("in the course of the training process, Defendants made hundreds of copies of copyrighted content")
Class Definition (¶ 294) "The Class consists of: All legal or beneficial owners of copyrighted works that: (A) were registered with the United States Copyright Office within five years of the work's first publication; (B) were downloaded or otherwise reproduced by OpenAI or Microsoft; (C) were registered with the United States Copyright Office before being downloaded or otherwise reproduced by OpenAI or Microsoft, or were registered within three months of first publication; and (D) are assigned one or more International Standard Books Number(s) (ISBN) or Amazon Standard Identification Number(s) (ASIN)."	*Alter*: "Whether Defendants' reproduction of the Class's copyrighted work constituted a copyright infringement." ¶ 115(a). *Tremblay*: "whether Defendants violated the copyrights of Plaintiffs and the Class when they downloaded copies of Plaintiffs' copyrighted books and used them to train ChatGPT." ¶ 58(a). *Chabon*: "Whether Defendants violated the copyrights of Plaintiffs and the Class when they downloaded and copied Plaintiffs' and the Class's copyrighted books." ¶ 60.
Commonality and Predominance (¶ 302) "Defendants have acted on grounds common to Plaintiffs and the Class by treating all Plaintiffs' and Class Members' works equally, in all material respects, in their reproducing and maintaining the Class's works without consent and or compensation, and in Defendants' LLM 'training.'"	*Alter*: "Whether Defendants' reproduction of the Class's copyrighted work constituted a copyright infringement." ¶ 115(a). *Authors Guild*: Microsoft built "the bespoke supercomputing system that OpenAI used to maintain and copy the copyrighted works." ¶ 424; *see also* ¶ 425 (Microsoft knew that "OpenAI was taking a massive corpus of copyrighted content, without compensation to rightsholders, and copying it for the purpose of training and developing its GPT models.").

	Tremblay: "[Novels] were copied into the BookCorpus dataset without consent, credit, or compensation to the authors. OpenAI also copied many books while training GPT-3." ¶¶ 39–40.

TABLE 3: COMPARING *BARTZ* AND *ALTER*	
Bartz (3:24-cv-05417 (N.D. Cal.), Dkt. 70)	*Alter* (No. 23-cv-10211 (S.D.N.Y.), Dkt. 26)
Headings "II. Anthropic Engaged in Largescale Copyright Theft in Training Its LLMs" "1. Large Language Models and the Training Process" "2. Anthropic Copied A Massive Trove of Pirated Books To Train Claude"	Headings "II. OpenAI and Microsoft Engaged in Largescale Copyright Infringement in Training the GPT Models" "1. GPT Models and the Training Process" "2. The Largescale Unlicensed Copying of Nonfiction Books to Train the GPT Models" "3. Defendants Reproduced Plaintiffs' Works To Train Their GPT Models"
Class Definition (¶ 63) "All natural persons, estates, literary trusts, and loan-out companies that are legal or beneficial owners of copyrighted works that: (a) are registered with the United States Copyright Office; (b) were or are used by Defendant in LLM training, research, or development, including but not limited to training Defendant's Claude family of models; and (c) are books."	Class Definition (¶ 109) "All owners of copyrighted literary works that: (a) are registered with the United States Copyright Office; (b) were or are used by Defendants in training their generative artificial intelligence models, including but not limited to GPT-3, GPT-3.5, GPT-4, and GPT-5; and (c) are works of nonfiction and have been assigned an International Standard Book Number (ISBN) and fall within a Book Industry Standards and Communications (BISAC) code other than Reference (REF).20."
Typicality (¶ 65) "The claims asserted by Plaintiff are typical of the claims of the Class, all of whose works were also copied as part of the LLM training process."	Typicality (¶ 111) "The claims asserted by Plaintiffs are typical of the claims of the Class, all of whose works were also copied as part of the GPT training process."

Commonality and Predominance (¶ 68(a))	Commonality and Predominance (¶ 115(a))
"Whether Anthropic's reproduction of the Class's copyrighted work constituted copyright infringement"	"Whether Defendants' reproduction of the Class's copyrighted work constituted copyright infringement"
Claim for relief (¶ 75)	Count I (¶ 121)
"Anthropic infringed on the exclusive rights, under 17 U.S.C. § 106, of Plaintiff and members of the proposed Class by, among other things, reproducing the works owed by Plaintiff and the proposed Class in datasets used to train their artificial intelligence models."	"Defendants infringed on the exclusive rights, under 17 U.S.C. § 106, of Plaintiffs and members of the proposed Class by, among other things, reproducing the works owed by Plaintiffs and the proposed Class in datasets used to train their artificial intelligence models."

TABLE 4: WRITTEN DISCOVERY	
RFPs	Second Set, Served on January 29, 2024 (Ex. 16) 23. DOCUMENTS and COMMUNICATIONS between YOU and MICROSOFT related to the process of collecting, maintaining, and using data to train CHATGPT, including the number of reproductions made of training data in the course of training CHATGPT. 30. All COMMUNICATIONS regarding Books1 and/or Books2, including, but not limited to, communications regarding the sources used to compile Books1 and/or Books2. 31. All DOCUMENTS sufficient to determine the source material for Books1 and/or Books2. Eighth Set, Served on January 9, 2025 (Ex. 17) 109. Documents sufficient to show all books You have downloaded and the source of such books. 110. Documents sufficient to show the dates You downloaded each book or dataset potentially containing books. 111. Documents sufficient to show any processes by which You cleaned, filtered or otherwise modified datasets potentially containing books prior to using them in Your large language models, including but not limited to any code employed to clean such datasets. 113. Documents Relating to Microsoft's belief or awareness of Your use or downloading of copyrighted material. 114. Documents Relating to Microsoft's belief or awareness that downloading copyrighted material or using it to train an LLM was not fair use. 124. Documents Relating to the "Books_corpus_reversible-sharded" or "Book-5B" datasets, including Documents sufficient to show their contents. Ninth Set, Served on January 10, 2025 (Ex. 18) 133. All Source Code related to LibGen, Books1, Books2, including, but not limited to, the code that downloaded "libgen-1-dedup" and "libgen-2-dedup" data from the LibGen websites and mirrors, code that discovered the list of books to download, scraped the webpages, and filtered and processed the data. REQUEST 134. All Source Code related to "ocr-books-internetarchive-202206-arrakis", including, but not limited to, the code that discovered the list of books to download, crawler code, OCR code, and any filtering and processing code.

	135. All Source Code related to Common Crawl and ELibra, including, but not limited to, the code that discovered the list of books to download, crawler code, OCR code, and any filtering and processing code. 136. All Source Code related to "podcasts-asr-en-20240102", including, but not limited to, the code that discovered the list of podcasts to download, crawler code, transcription code, and any filtering and processing code. 137. All Source Code related to downloading YouTube videos, including, but not limited to, code on transcribing and any arrangement with Microsoft or Google. Tenth Set, Served on January 23, 2025 (Ex. 19) 147. Documents regarding reports provided by any of the Start-Up Fund Entities to the other OpenAI Defendants regarding any OpenAI LLM, the use or acquisition of books for training of any LLM, or copyright issues concerning LLMs. 149. Documents or Communications regarding the use or acquisition of books for training any LLM between any Start-Up Fund Entity and OpenAI OpCo, LLC. 151. Documents sufficient to identify any investments made by any Start-Up Fund Entity related to (1) any company that has products or services involved in the training of any OpenAI LLM, (2) the use or acquisition of books for training of an LLM, (3) copyright issues concerning LLMs, or (4) any company that relies upon OpenAI LLMs.
ROGs	First Set, Served on February 23, 2024 (Ex. 20) 1. Identify the five OpenAI Individuals who are most knowledgeable about the creation and contents of the Books2 dataset. 2. Identify the five OpenAI Individuals who are most knowledgeable about Your data retention policies. Third Set (in *Tremblay*), Served on March 14, 2025 (Ex. 9) 21. Identify all datasets in your possession that contain copyrighted works. 22. Identify all datasets in your possession that contain the Asserted Works. 23. For all datasets identified in response to Interrogatories 21 or 22 describe how the data was obtained. 24. For all the datasets identified in response to Interrogatories 21 or 22 describe whether the datasets, or data derived from the datasets, were shared with or made available to any third party. For each instance of such sharing, specify: a. The identi[t]y of the third-party,

	b. How much data was shared, c. Whether the data shared included copyrighted works, d. Whether the data included the Asserted Works, and e. The mechanism of sharing, i.e., a data-sharing agreement, seeding through a torrent, sale or trade of data. 25. For all datasets identified in response to Interrogatories 21 or 22 describe how the dataset was used by OpenAI, including specifying whether the datasets were used in development, training, validation, testing, filtering, evaluation, research, analysis, application development, or any other purpose related to large language models and specifying which language model the datasets were used in relation to; if the dataset was not used by OpenAI describe why it was not used and specify what happened to the dataset. [Definition] 3. "Dataset(s)" means a corpus, collection, file, directory or folder, of material including or potentially including copyrighted material, including but not limited to the type of databases described in paragraph 44 of the Complaint – databases including but not limited to BooksCorpus, Books3, Z-Library (aka B-ok), Library Genesis (aka LibGen), Bibliotik, Anna's Archive, Internet Archive, and the Pile, or any other database or compiled sets of documents or data. The term applies to any such corpus, collection, file, directory or folder, regardless of whether it was acquired as a discrete collection or treated as such by You, that was used, referenced, considered, or intended to be used in connection with the development, training, validation, testing, or evaluation, research, analysis, application development, or any other purpose related to large language models, even if not ultimately used for training or having any influence or effect on the same, and extends to all versions, updates, augmentations, or modifications of any large language model. The term encompasses all versions, updates, augmentations, or modifications of such datasets. Second Set, Served on April 5, 2024 (Ex. 21) 8. Identify all Non-OpenAI Individuals to whom You gave access to Books1, or any portion thereof, or who otherwise possessed or possess a copy of Books1, or any portion thereof. 9. Identify all Non-OpenAI Individuals to whom You gave access to Books2, or any portion thereof, or who otherwise possessed or possess a copy of Books2, or any portion thereof. Revised Third Set (in *Tremblay*), Served on June 4, 2025 (Ex. 10) 21–25 [revised]: "Nonetheless, as a compromise, we would agree to accept a list of the names of each dataset that OpenAI downloaded from a known shadow library,

	listed below, along with the date and method of download, and production of those full datasets: Library Genesis ("LibGen") Anna's Archive Pirate Library Mirror ("PiLiMi") Sci Hub Z-Library Bibliotik / Books3 UbuWeb The Pirate Bay Interplanetary File System ("IPFS")"

CERTIFICATE OF SERVICE

I hereby certify that on August 14, 2025, I caused the foregoing document to be electronically filed with the Clerk of the United States District Court for the Southern District of New York using the CM/ECF system, which shall send electronic notification to all counsel of record.

/s/ Henry Walter
Henry Walter

CERTIFICATE OF WORD COUNT

Pursuant to Local Civil Rule 7.1, I hereby certify that the foregoing document was typed using 12-point, Times New Roman font and contains 7,163 words, exclusive of the case caption, table of content, table of authorities, signature blocks, appendix, and certificates, but does include material in footnotes.

/s/ Henry Walter
Henry Walter